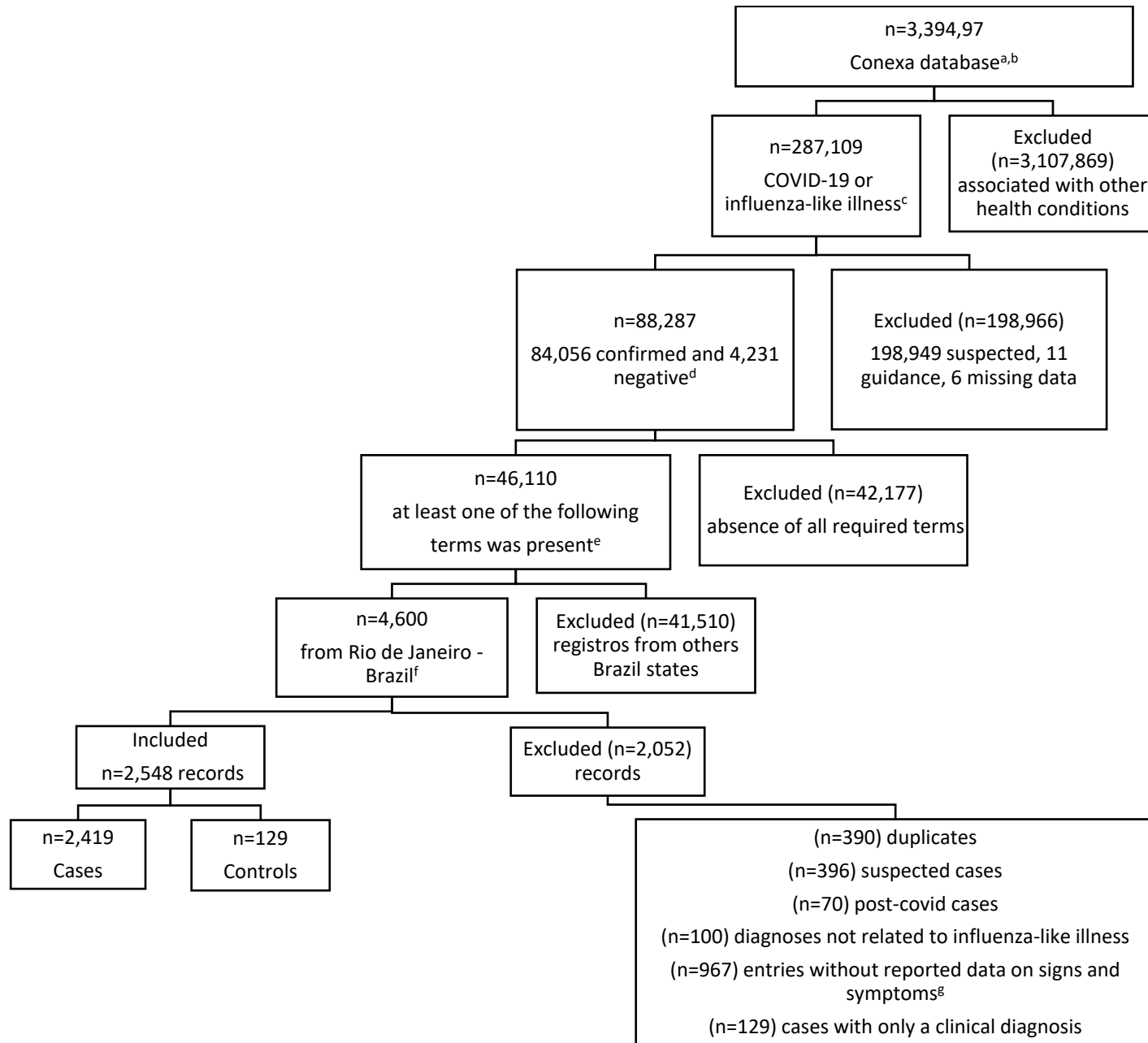


Figure S1: Flowchart of data selection for the development of a diagnostic prediction model for COVID-19.



- a. Variables used: *Date of Care; Care Completion Date; Patient ID; Patient State (UF); Epidemic; Epidemic (Medical Report); Sex; Age; Patient Report*.
- b. Selection of records related to the topic (COVID-19 or flu-like syndrome) was performed using the variable 'epidemic'.
- c. Selection of records classified as confirmed or negative was performed using the variable 'epidemic (medical report)'.
- d. Selection using an Excel function [=IF(ISNUMBER(SEARCH(searched_text; selected_variable)); "1"; "0")] applied to the variable '*patient report*'. Keywords used in the function: *positive, negative, detectable, detected, confirmed, test, exam, covid, corona*.
- e. Selection by place of origin.
- f. The variable '*patient report*' was analyzed across all 4,660 records to exclude duplicates, select cases and controls, and identify reported signs and symptoms.
- g. Records without specification of signs and symptoms were classified as missing data.

Table S1. Sex and Age observed between COVID-19 and control cases

	Cases (n=2289)	Controls (n=122)	Total (n=2411)	p-value
	n (%)	n (%)	n (%)	
Sex at birth ^a Woman	1381 (60.3)	81 (66.4)	1462 (60.6)	0,182 ^c
Man	908 (39.7)	41 (33.36)	949 (39.4)	
	Mean (SD)	Mean (SD)	Mean (SD)	
Age ^d	44.2 (14.8)	40.8 (16.0)	44 (15)	0.008 ^d

n: absolute number; %: percentagem value; SD: standard deviation; ^a137 missing data; ^b5 missing data; ^cquiquadrado test; ^dt test